BeBeC-2024-D09

THREE-DIMENSIONAL POSITION ESTIMATION OF SOUND SOURCES WITH MICROPHONE ARRAYS

Bence Csóka[1], Péter Fiala[1] and Péter Rucz[1]
[1]Department of Networked Systems and Services, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

ABSTRACT

This paper is concerned with the three-dimensional position estimation and tracking of sound sources with microphone arrays and acoustical beamforming. Beamforming algorithms are effective methods for mapping a sound field. The position estimation based on beamforming usually only covers the direction of the sound source, but it can be extended to include distance estimation as well.

The basis of our 3D position estimation algorithm is the Delay-and-Sum method, and we enhance it with more advanced beamforming algorithms, like the MUSIC algorithm. The distance estimation is performed by extending the two-dimensional observation plane into three dimensions. We also use the Kalman-filter algorithm to properly track moving sound sources, instead of just taking a series of snapshots of them.

Our goal is to evaluate the 3D position estimation method through simulations, outdoor measurements of Unmanned Aerial Vehicles (UAVs) and indoor measurements performed in a semi-anechoic chamber. Simple simulations with ideal environmental conditions yielded promising results, but the algorithm is only reliable for outdoor measurements when direction estimation is performed without distance estimation. We propose improvements on the 3D estimation process to increase its accuracy and reliability, to enable its application in real-world scenarios.

1 INTRODUCTION

Beamforming is a widely used method for the localization of objects in many different areas. In the case of objects emitting sound, acoustical beamforming is a viable method for sound source localization. Our main goal is to estimate the position of moving sound sources with microphone arrays used as acoustical cameras. Standard Delay-and-Sum Beamforming in the frequency domain is enhanced with more advanced beamforming algorithms, such as Multiple Signal Classification (MUSIC), and Iterative Sparse Asymptotic Minimum Variance (SAMV) for higher resolution sound maps. Both MUSIC and SAMV have been well discussed in the scientific literature, with several different variations for different situations

and purposes. Xenaki, Gerstoft and Mosegaard compare conventional beamforming and MUSIC with more modern methods [1]. Gupta and Kar developed an improved version of MUSIC that is capable of mapping coherent sources [2]. Yaning, Juntao and Xinghao and Le devised a version of the algorithm with decreased computational complexity [3]. The SAMV algorithm also has different variants [4,5] that work better either in low or high SNR conditions.

The position estimation with beamforming usually only means direction estimation, but it can be extended into three dimensions to include distance estimation as well. This is a relatively novel concept in the field of acoustics, but there have been a few research initiatives similar to ours. Cai and co. combined beamforming with a binocular camera for the purpose of three-dimensional sound field reconstruction [6]. Valin and co. developed a 3D localization method for video conferences that worked in the near-field, up to 3 meters [7]. In 2022, Merino-Martinez presented a distance estimation method where asynchronous measurement data from the same microphone array at multiple locations was used for quasi-stationary sound sources [8]. Also in 2022, Sarradj presented 3D source mapping with gridless orthogonal beamforming with improved resolution and decreased computational cost compared to methods using discrete grids [9]. Liaquat and co. devised a 3D localization method for microphone arrays consisting of a low number of sensors [10]. In contrast, our aim is to develop a purely acoustical method employing 3D beamforming with a discrete grid, to localize moving sound sources, with and array consisting of 48 microphones.

We will first discuss the methodology of our approach, the basics of beamforming, the Delay-and-Sum method, and the more advanced MUSIC and SAMV algorithms. Also, as part of the methodology, we introduce our extension of the beamforming grid to facilitate 3D position estimation. We also cover the integration of the Kalman-filter algorithm into the process to predictively track moving sources instead of just taking momentary snapshots of the sound field. Next, we present preliminary simulation and measurement results using MUSIC. Finally, we discuss present and potential improvements on the created algorithm, such as using the more robust beamforming method SAMV, and adaptively fitting the observed frequency on peaks in temporally changing frequency spectra.

2 METHODOLOGY

2.1 Focusing and source localization

There are two main tasks that must be performed in our approach for position estimation: focusing and source localization. Focusing is the enhancement of sound arriving from a specific direction and the suppression of sound arriving from other directions. It is based on the Delay-and-Sum method, which is the appropriate steering and then superposition of the received sound signals of the microphones. The steering of the signals consists of amplification and delay, and its purpose is to counteract the amplitude and phase differences due to propagation between the source and the sensors in different positions (Fig. 1.). This results in an amplified superimposed signal when focusing on the direction of a sound source, and attenuation when focusing on directions where no source is present [11].

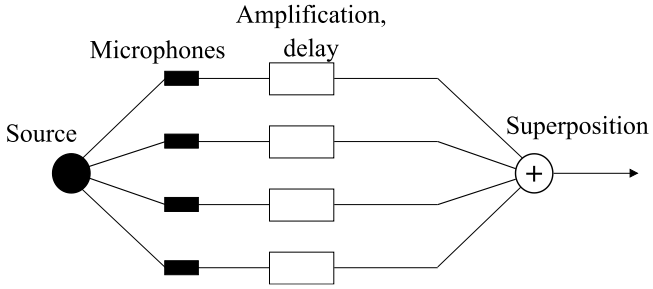


Fig. 1. Acoustical focusing with the Delay-and-Sum method: the appropriate amplification, delay, and then superposition of the received signals of the microphones to enhance sound arriving from a specific direction, while other directions are suppressed.

Source localization is the estimation of the position of the sound source on a set of points in space, which is called acoustical canvas or scanning grid. The points of the canvas are all treated as potential source positions, and the likelihood of a source being present at each one is assessed. Virtual sound sources are placed on these points one by one, and the more similar the generated virtual sound field is to the real one, the higher the aforementioned likelihood is. Finally, the point of the grid with the highest likelihood is the estimated position of the sound source.

These two tasks can be performed in conjunction with one another by following these steps:

1. We focus on one of the points of the scanning grid with the Delay-and-Sum method for the duration of a short time window.
2. We take the frequency spectrum of the focused signal and observe a narrow band from it around a chosen frequency.
3. We consider the energy of this narrow band as a representation of the likelihood of a sound source being present in that position, that emits energy at least in the observed frequency range: the higher the energy, the higher the likelihood.
4. The first three steps are repeated for every point of the scanning grid. Different colours are assigned to different likelihoods, expediently warmer colours to higher likelihoods, and a sound map is drawn with the use of these colours. This map is an easily interpretable visual representation of the estimated sound field.
5. Source localization on this sound map is equal to looking for local likelihood maxima, or spots with warmer colour than their surroundings.
6. The first five steps can be repeated for different time windows, thus creating snapshots of the sound field at different points in time. This opens the possibility of observing temporal changes in the sound field, for example in the case of moving sound sources.

This algorithm based on the Delay-and-Sum method is a basic and simple solution to the problem of source localization. While one microphone has a uniform directionality, the array as a whole has a non-uniform directionality that can be steered together with the received signals. This directional characteristic can also be enhanced by employing more advanced beamforming methods (such as MUSIC and SAMV) to achieve a narrower main lobe and more supressed sidelobes.

2.2 Steering vectors and cross-spectral matrix

Beamforming is used to estimate the source distribution vector ($\boldsymbol{x}$), the elements of which correspond to the points of the acoustical canvas. The true source distribution is

unknown, and its estimation is performed with the known vector of the received signals (***y***) and the steering matrix (***A***). Both ***x*** and ***y*** contain information in a narrow frequency band (as it was mentioned in Section 2.1), and the elements of ***y*** correspond to the microphones. The steering matrix consists of the steering vectors between the scanning grid and the sensors, and it gives the propagation information in the form of complex numbers:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{1}$$

$$\boldsymbol{A}(i,j) = e^{jkd_{i,j}} * d_{i,j}. \tag{2}$$

The propagation depends on the distance between the i-th microphone and the j-th point of the canvas ($d_{i,j}$). The steering values are normalized with the square root of the number of microphones (M). This way, the steering values are the inverse of the amplitude decrease and phase shift due to propagation, and thus serve as compensation for them.

A very important concept for several beamforming algorithms is the cross-spectral matrix (CSM). In this case, the CSM gives the spectral cross-correlation between the received signals of the microphones. It can be defined with the help of unknown signal powers and noise variance:

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H + \sigma\boldsymbol{I}, \tag{3}$$

where ***P*** is a diagonal matrix containing the source strengths on the main diagonal, σ is the noise variance, and H denotes the Hermitian transpose. The CSM can be estimated with the received signals:

$$\boldsymbol{R}_N = \boldsymbol{Y}\boldsymbol{Y}^H/N, \tag{4}$$

where N is the number of snapshots used for the estimation, and **Y** is the matrix containing the **y** vectors for the corresponding time windows.

2.3 MUSIC algorithm

The MUSIC algorithm is a simple linear algebraic method that is based on the eigenvalue-decomposition of the CSM:

$$\boldsymbol{R}_N = \boldsymbol{U}\sum\boldsymbol{U}^H, \tag{5}$$

where ***U*** is a unitary matrix, whose columns are the eigenvectors, and ∑ is a diagonal matrix whose diagonal elements are the eigenvalues. In the traditional version of the MUSIC algorithm, the number of sound sources (K) is estimated in advance, and the eigenvectors corresponding with the K largest eigenvalues make up the signal subspace, while the rest make up the noise subspace ($\boldsymbol{U_n}$). This noise subspace and the steering matrix are then used for calculating the sound map [1,2,12,13,14]:

$$\boldsymbol{P_{MUSIC}} = \frac{1}{\boldsymbol{A^H U_n U_n^H A}}, \tag{6}$$

which means, that the elements of $\boldsymbol{P_{MUSIC}}$, the estimation of the source distribution, are the reciprocal of the elements of the $\mathbf{A^H U_n U_n^H A}$ matrix. The main advantage of this algorithm compared to conventional beamforming methods is its better noise tolerance and higher resolution, without significantly increased computational cost. However, it only works for uncorrelated sound sources.

2.4 SAMV algorithm

The SAMV algorithm is an iterative method using the CSM of the received signals to create a sparse sound map. The sparsity of the sound map means that out of all the energies/likelihoods assigned to the points of the scanning grid, only a select few are nonzero, while all the others are small enough to be negligible. During each iteration, the CSM, the estimated source strengths and the estimated noise variance are updated. The biggest advantage of this method compared to MUSIC is that it can be used to localize correlated sources, and a higher resolution, but with an increased computational cost.

SAMV is based on the asymptotically minimum variance (AMV) approach [4]. The goal of the AMV approach is to estimate the source distribution ($\hat{\boldsymbol{p}}$) by minimizing the following criterion:

$$\hat{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}} f(\boldsymbol{p}), \tag{7}$$

$$f(\boldsymbol{p}) \stackrel{\text{def}}{=} [\boldsymbol{r}_N - \boldsymbol{r}(\boldsymbol{p})]^H \boldsymbol{C}_r^{-1} [\boldsymbol{r}_N - \boldsymbol{r}(\boldsymbol{p})], \tag{8}$$

where $\boldsymbol{C}_r$ is the following Kronecker-product:

$$\boldsymbol{C}_r = \boldsymbol{R}^* \otimes \boldsymbol{R}. \tag{9}$$

The * denotes the conjugate transposed of a matrix, and $\boldsymbol{r}_N$ and $\boldsymbol{r}(\boldsymbol{p})$ are the vectorization of $\boldsymbol{R}_N$ and $\boldsymbol{R}$, respectively. This means that we are looking for an estimated source distribution that minimizes the difference between the measured and estimated covariance. SAMV achieves this through an iterative process and gives a sparse solution for the source distribution.

The first step of the SAMV algorithm is the initialization of the source strengths (p_k) for every point of the scanning grid, and the noise variance (σ) [4,5]:

$$p_k^{(0)} = \frac{\boldsymbol{a_k^H R_N a_k}}{\left||a_k|\right|^4}, \tag{10}$$

$$\sigma^{(0)} = \frac{1}{MN} \sum_{n=1}^{N} \left||\boldsymbol{y}(n)|\right|^2, \tag{11}$$

where $\boldsymbol{a_k}$ is the steering vector for the k-th point of the grid, M is the number of microphones, and N is the number of snapshots used to determine $\boldsymbol{R_N}$ in equation (4).

After the initialization, the iteration begins:

1. First, the CSM ($\boldsymbol{R}$) is updated with equation (3).
2. Second, the source strengths are updated. There are different versions of the SAMV algorithm with different formulas for this step, and for our research, the SAMV-2 approach from [4] was chosen:

$$p_k^{(i)} = \frac{\boldsymbol{a_k^H R^{-1(i)} R_N R^{-1(i)} a_k}}{\boldsymbol{a_k^H R^{-1(i)} a_k}}. \tag{12}$$

3. Finally, the noise variance is updated:

$$\sigma^{(i)} = \frac{Tr(\boldsymbol{R^{-2(i)} R_N})}{Tr(\boldsymbol{R^{-2(i)}})}, \tag{13}$$

where Tr() denotes the trace of the matrix.

These three steps are repeated an arbitrary number of times. The number of iterations is chosen considering a good compromise between computational time and the sparsity of the sound map.

2.5 Distance estimation

Source localization usually only covers direction estimation, but distance estimation can also be included by extending the two-dimensional scanning grid into three dimensions. This is based on the dependence of the quality of the sound map on the difference between the source distance and the focal distance. The closer the focal distance is to the sound source, the better the map becomes, with the main beam width of beamforming becoming narrower. Fig. 2. shows the result of a simple simulation where a point source is localized by the MUSIC algorithm. The source is 5 meters from the microphone array, and the focal distance is either 1, 3, 5, 10 or 100 meters. The focal distance has a significant impact on the sound map, and when it equals 5 meters, MUSIC gives a sparse solution. This phenomenon can be used to our advantage: distance estimation can be achieved with a 3D acoustical canvas that consists of points at many different distances from the array.

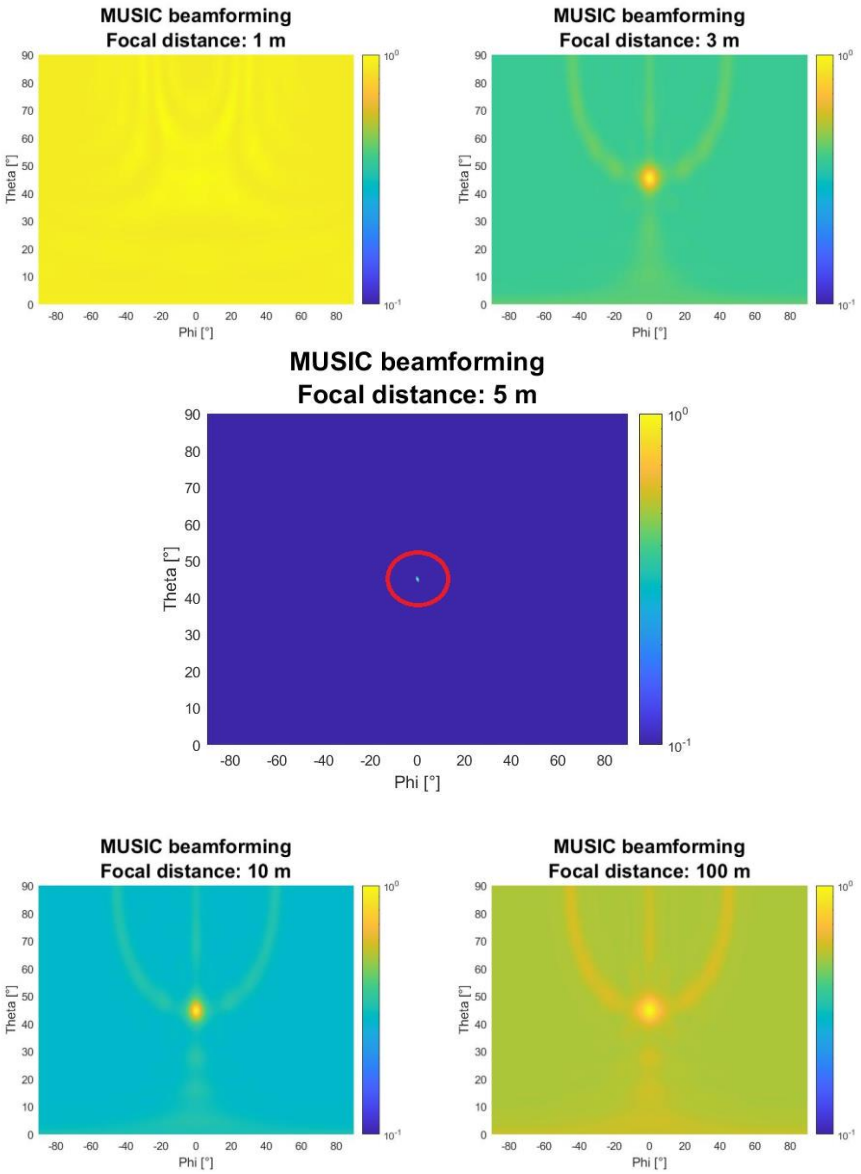


Fig. 2. Sound map of the same source distribution with different focal distances. The estimations of MUSIC are divided by their maxima, and their logarithms are plotted as a two-dimensional function of the direction.

One possible approach to extend the canvas into 3D is to make a direction estimation the usual way on a primary canvas, and to create a secondary canvas in the estimated direction (Fig. 3.). This secondary canvas is a discretized line that consists of many points at many different distances, but they are all in the same direction. Beamforming is applied on this secondary canvas, and the maximum corresponds with the estimated position of the source. This approach is more computationally efficient than using a fully three-dimensional grid. Fig. 4. shows a typical result of beamforming on the secondary canvas, which consists of points placed densely between 0.01 and 1000 meters in a partially logarithmic manner. In the two simulations, the two stationary sources are at 5 and 50 meters, and the maxima of the beamforming on the secondary canvas are 4.99 meters and 49.7 meters, respectively, which

means, that these are the estimated distances. This approach works both with MUSIC and SAMV.

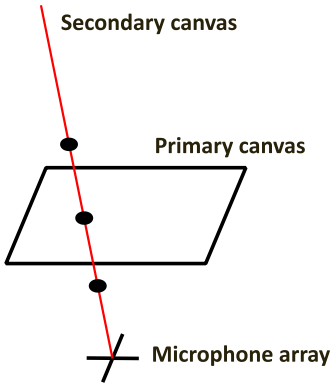


Fig. 3. Extending the acoustical canvas into three dimensions by creating a secondary canvas in the initially estimated direction.

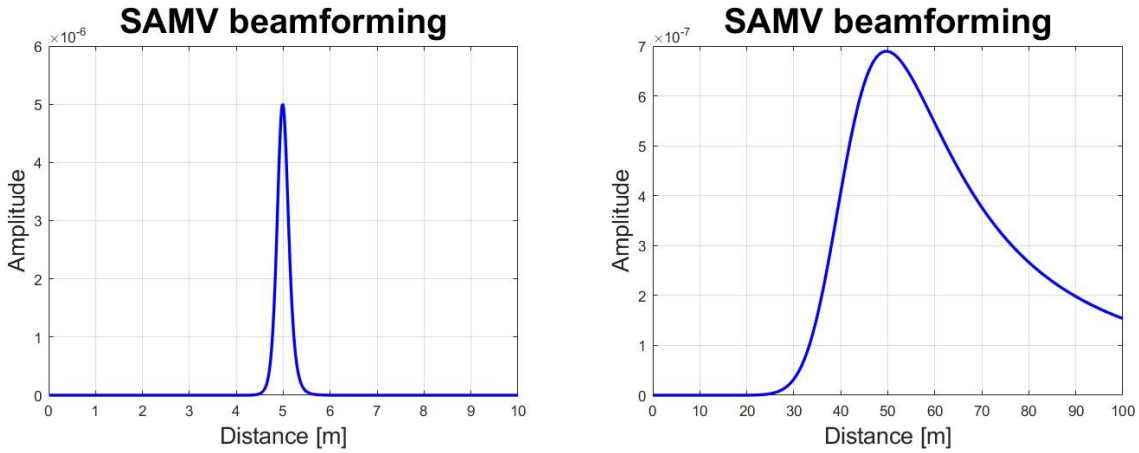


Fig. 4. Distance estimation by applying beamforming on the secondary canvas. One of the sources is 5 meters from the microphone array (left), the other is at 50 meters.

2.6 Kalman-filter

So far, the methods discussed here can create momentary snapshots of the sound field, but they can be extended with the Kalman-filter algorithm for the predictive tracking of sound sources. The Kalman-filter gives an optimal estimation of the state of temporally dynamic systems [15]. In this case, the system in question is a moving sound source, and its state can be defined as its position and velocity. The algorithm considers the measurement data (which is the output of beamforming as position coordinates, either in 2D or 3D), and on top of that, the earlier states of the system. This means, that the position estimation is based on more information than when only beamforming algorithms are used, which results in higher accuracy, provided the parameters of the algorithm are tuned properly. Using the Kalman-filter also opens the possibility of predicting and tracking the movement trajectory of the observed object.

Assuming that the system is linear and time-invariant, we can start with the standard discrete state equation:

$$\boldsymbol{x}(n+1) = \boldsymbol{A}\boldsymbol{x}(n) + \boldsymbol{B}\boldsymbol{u}(n) + \boldsymbol{w}(n). \tag{14}$$

$$\boldsymbol{y}(n) = \boldsymbol{C}\boldsymbol{x}(n) + \boldsymbol{D}\boldsymbol{u}(n) + \boldsymbol{v}(n). \tag{15}$$

Here ***x(n)*** is the state vector during the *n*-th snapshot / time window. In three dimensions, this vector consists of six elements, which are the three position and the three velocity coordinates. ***u(n)*** is the input excitation vector, and ***y(n)*** is the output vector. ***A***, ***B***, ***C*** and ***D*** are system matrices (***D*** is negligible because the input doesn't have a direct impact on the output). ***w(n)*** and ***v(n)*** are the process noise and measurement noise vectors respectively, representing the inaccuracies of the model and the measurements. These two noise vectors are uncorrelated, normally distributed with zero mean and covariance matrices denote with ***Q(n)*** and ***R(n)***.

The first step is an a-priori estimation (denoted with a "-" upper index) of the state and output vectors:

$$\boldsymbol{x}^{-} = \boldsymbol{A}\tilde{\boldsymbol{x}}(n) + \boldsymbol{B}\boldsymbol{u}(n), \tag{16}$$

$$\tilde{\boldsymbol{y}}(n) = \boldsymbol{C}\boldsymbol{x}^{-}(n). \tag{17}$$

The difference between the measurement ($\boldsymbol{y}(n)$) and the estimation ($\tilde{\boldsymbol{y}}(n)$):

$$\boldsymbol{d}(n) = \boldsymbol{y}(n) - \tilde{\boldsymbol{y}}(n). \tag{18}$$

This difference is then used for an a-posteriori estimation (denoted with an upper "+" index):

$$\tilde{\boldsymbol{x}}(n+1) = \boldsymbol{x}^{+} = \boldsymbol{x}^{-} + \boldsymbol{K_n}\boldsymbol{d}(n), \tag{19}$$

where $\boldsymbol{K_n}$ is a correction matrix. The optimal correction matrix is found with equations (20)-(22):

$$\boldsymbol{P_n^{-}} = \boldsymbol{A}\boldsymbol{P_{n-1}}\boldsymbol{A}^T + \boldsymbol{Q_n}, \tag{20}$$

$$\begin{aligned} \boldsymbol{P_n^{+}} &= (\boldsymbol{I} - \boldsymbol{K_n}\boldsymbol{C})\boldsymbol{P_n^{-1}}(\boldsymbol{I} - \boldsymbol{K_n}\boldsymbol{C})^T + \boldsymbol{K_n}\boldsymbol{R_n}\boldsymbol{K_n^T} = \\ &= (\boldsymbol{P_n^{-1}} + \boldsymbol{C}^T\boldsymbol{R_n^{-1}}\boldsymbol{C})^{-1} = \\ &= (\boldsymbol{I} - \boldsymbol{K_n}\boldsymbol{C})\boldsymbol{P_n^{-}}, \end{aligned} \tag{21}$$

$$\boldsymbol{K_n = P_n^- C^T (C P_n^- C^T + R_n)^{-1} = P_n^+ C^T R_n^{-1}}, \tag{22}$$

where $\boldsymbol{P_n^-}$ and $\boldsymbol{P_n^+}$ are the covariance matrices of the a-priori and a-posteriori state vectors, respectively.

This traditional version of Kalman-filter can estimate the state of linear systems, but in real life situations, the observed system is often nonlinear. In our implementation, the Kalman-filter gets receives the measurement data as spherical coordinates, so the algorithm must be extended to handle nonlinear systems as well. One such extension is the Unscented Kalman Filter (UKF) algorithm [16].

UKF creates $2N_d$ sigma points around the state vector for every snapshot, where N_d is the number of dimensions in the state space:

$$\boldsymbol{x_i^\sigma, x_{N_d+i}^\sigma = x_n \pm \sigma_i}, \qquad i = 1 \dots N_d, \tag{23}$$

where $\boldsymbol{\sigma_i}$ is the i-th row of the $\sqrt{\boldsymbol{NP_n}}$ matrix. Because the sigma points were defined this way, their statistical average and variance are equal to the state vector and its covariance matrix. Next, equation (16) is applied on the sigma points, and the resulting points are denoted with $\boldsymbol{x_i^{\sigma *}}$. The a-priori state vector and its covariance are then calculated as follows:

$$\boldsymbol{\tilde{x}^-} = \frac{1}{2N} \sum_{i=1}^{2N} \boldsymbol{x_i^{\sigma *}}, \tag{24}$$

$$\boldsymbol{\tilde{P}^-} = \left(\frac{1}{2N} \sum_{i=1}^{2N} (\boldsymbol{x_i^{\sigma *}} - \boldsymbol{\tilde{x}^-}) \, (\boldsymbol{x_i^{\sigma *}} - \boldsymbol{\tilde{x}^-})^T \right) + \boldsymbol{Q}. \tag{25}$$

Then, new sigma points are created with these parameters similarly to equation (23), and the (17) output equation is applied on these new sigma points. The average of the $\boldsymbol{y_i^\sigma}$ resulting points is denoted with $\boldsymbol{\tilde{y}}$. The auto- and cross-correlation matrices are determined with equations (26) and (27):

$$\boldsymbol{P_{yy}} = \frac{1}{2N} \sum_{i=1}^{2N} (\boldsymbol{y_i^\sigma} - \boldsymbol{\tilde{y}})(\boldsymbol{y_i^\sigma} - \boldsymbol{\tilde{y}})^T, \tag{26}$$

$$\boldsymbol{P_{xy}} = \frac{1}{2N} \sum_{i=1}^{2N} (\boldsymbol{x_i^{\sigma *}} - \boldsymbol{\tilde{x}^-})(\boldsymbol{y_i^\sigma} - \boldsymbol{\tilde{y}})^T. \tag{27}$$

The correction matrix is derived from these correlation matrices:

$$\boldsymbol{K_n = P_{xy} P_{yy}^{-1}}. \tag{28}$$

Finally, the correction matrix is used in the a-posteriori estimation:

$$\boldsymbol{x_{n+1} = \tilde{x}^{+} = x^{-} + K_n(y_n - \tilde{y})}, \tag{29}$$

$$\boldsymbol{P_{n+1} = P^{+} = P^{-} + K_n\left(P_{yy} + R\right)K_n^T}. \tag{30}$$

3 SIMULATION AND MEASUREMENT RESULTS

3.1 Preliminary simulation with MUSIC

In this section, we present a simple simulation example with ideal conditions, where the sound source localization (both direction and distance estimation) is performed with the MUSIC and Kalman-filter algorithms. The simulation was run in the MATLAB environment.

The simulated microphone array consists of 48 sensors placed in a cross formation, and the distance between adjacent microphones is 6 centimetres, and thus the upper frequency limit for the spatial overlap is slightly above 2.8 kHz. The primary canvas consists of 20000 (200 times 100) points evenly distributed on a rectangular area 15 meters from the microphone array (Fig. 5.). The secondary canvas always changes: it lies in the direction estimated on the primary canvas, and consists of 4500 points, whose distances from the centre of the array are distributed in a partially logarithmic fashion between 0.01 and 1000 meters. One point source is emitting filtered white noise and is moving with constant velocity parallel to the array plane. In three different simulations it is either 5, 25 or 50 meters from the microphones, and its velocity is 1, 5 or 10 m/s, respectively. This way, the velocity is small enough that the Doppler-effect is negligible. The sound-to-noise ratio (SNR) is 10 dB (that is, the ratio of the variances of the "useful" and background white noises) and the time windows are 50 milliseconds long.

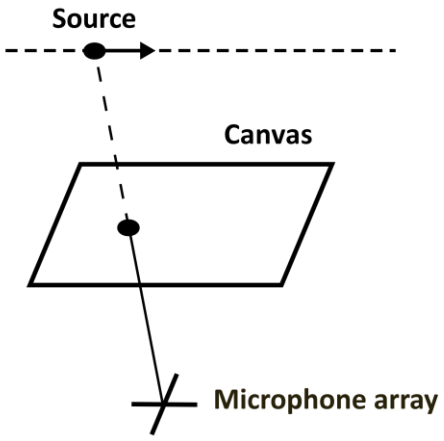
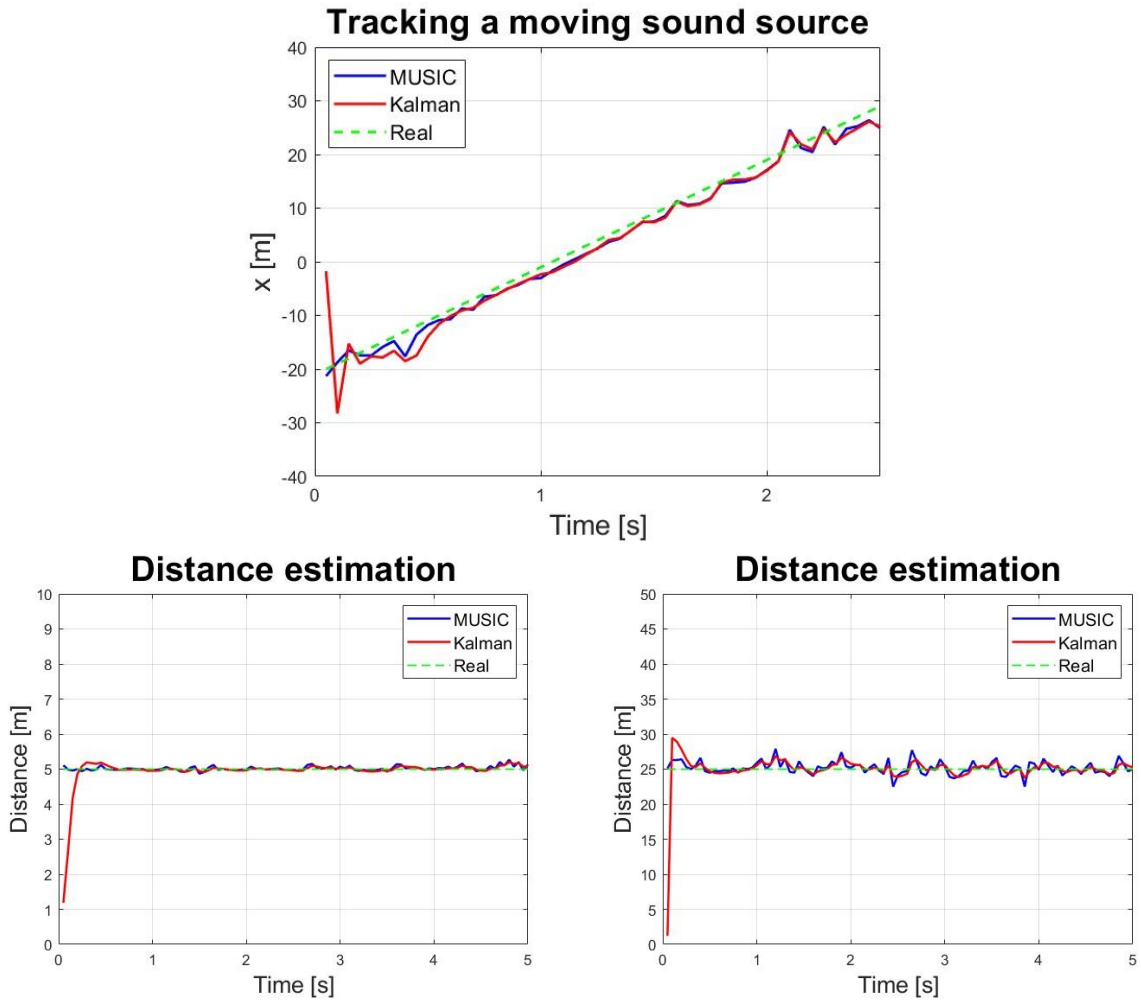


Fig. 5. Simulation arrangement: microphones are in a cross formation, the primary canvas on a rectangular area parallel to the array, and a sound source performing uniform motion parallel to both.

Direction estimation with MUSIC is successful in this simulation example, and Kalman further reduces the mean square error of the estimation. Distance estimation proves to be more challenging due to greater relative variance around the beamforming maxima, but it is still successful (Fig. 6.). It is more accurate for closer sources, and the variance of the estimated distance is greater relative to the actual distance for farther sources. This is because when the size of the array becomes negligible compared to the source distance, the wave propagation is closer to planar, and slight changes in the distance result only in small changes in the angles of incidence. Nevertheless, this example is a promising starting point for the fully three-dimensional position estimation algorithm, at least for ideal environmental conditions.



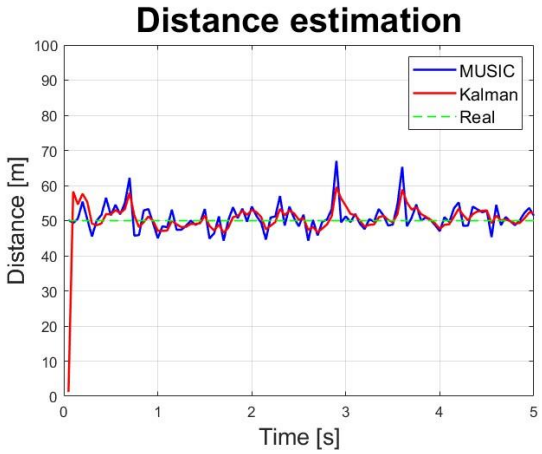


Fig. 6. Direction and distance estimation of a moving sound source with the MUSIC and Kalman-filter algorithms, with a sound source being at 5, 25, or 50 meters from the sensor array.

3.2 Preliminary measurement with MUSIC

Direction and distance estimation was successful in a simple simulation example. However, the end goal is for the algorithm to be applicable in real-life situations, so it is important to test it in less favourable conditions.

During our work, we participated in outdoor measurements where unmanned aerial vehicles (UAVs), or drones served as sound sources. The measurements presented here are of two drones named Secopx8 and Tarot680. The microphone arrangement used here is the same as in the previous simulation (48 microphones in a cross formation), and so is the primary scanning grid (evenly distributed points on a rectangular area). The microphones are stuck firmly in appropriately sized holes in a wooden board, and this board is positioned close to upright. There is a webcamera placed on the top of the board to provide a video recording of the flying drones that can be fitted onto the sound maps created with MUSIC.

Fig. 7. shows the estimation of MUSIC and Kalman-filter for one snapshot of each of the drones. In both cases, the UAV is flying in front of the array with a couple of meters of distance. For most time windows, direction estimation is successful, except for a few moments when the determined position is incorrect, presumably due to a strong background noise or ground reflection. Distance estimation, however, is unsuccessful due to far from ideal environmental conditions. The output of the algorithm changes too erratically with each new snapshot, which can't be an accurate reflection of reality, because the drone moves slowly.

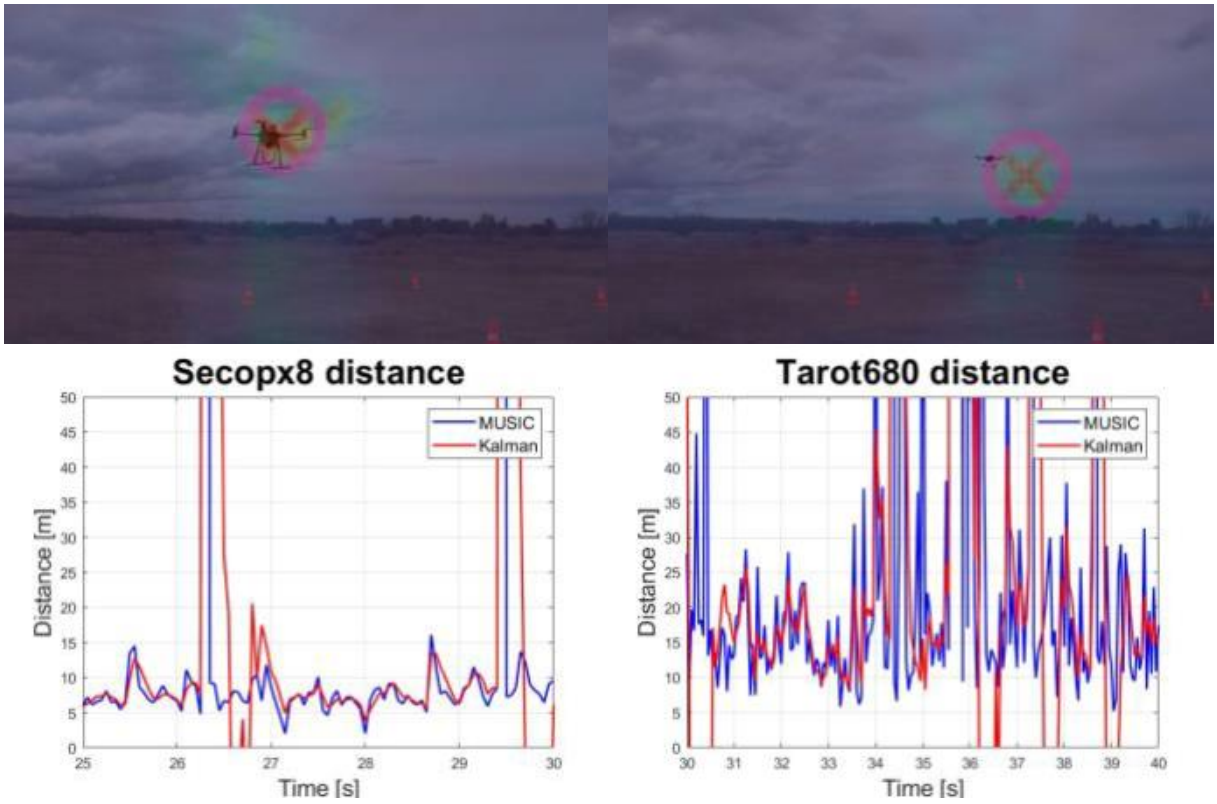


Fig. 7. Direction and distance estimation of Secopx8 (left) and Tarot680 (right).

3.3 Comparison of preliminary results, discussion

The main takeaway from the preliminary results is that direction estimation with the MUSIC and Kalman-filter algorithms is successful, both in simulations, and in measurements with good enough environmental conditions. Unfortunately, distance estimation was only possible during an idealized, simplified simulation. To achieve applicability for real-life situations, it is important to consider and investigate unfavourable environmental conditions, and to account for them in the algorithm.

As the simulation was only a simplification of reality, many factors were neglected that potentially can be critical in preventing successful distance estimation in measurements. Potential critical differences include:

- The finite extent of the sound source. In the simulation, the source was modelled with a point source, while drones (and other sound sources) have an extent that is not negligible when they are close to the array.
- The emitted sound. In the simulation, the source emitted filtered white noise, and the observed frequency range didn't change. However, in real life, temporal changes in the frequency spectrum must be accounted for by adaptively changing the observed narrow band.
- The directivity of the source, which was assumed to be uniform in the simulation, but it's rarely the case for drones.
- The background noise is usually much more irregular than the simulated white noise.
- The presence of ground reflections, which was neglected during the simulation.

- The trajectory and velocity of the moving source, which is much more irregular compared to uniform motion.

So far, out of these factors, we started investigating ground reflections, and the relation between the emitted sound and the observed frequency band.

3.4 Ground reflections

One of the main advantages of the SAMV algorithm over MUSIC is that it can localize correlated sources. This makes it a useful approach in cases when ground reflections are present. This section compares the performance of the two algorithms through measurements performed in a semi-anechoic chamber, where both the presence and absence of ground reflections can be set up. The microphone array is the same 48 channel cross formation as before, and the primary acoustical canvas is on a rectangular area 5 meters from the array. A stationary cell phone serves as the sound source that emits either a generated harmonic signal, or a recording of the sound of a UAV. The time windows for processing the received signals are 0.1 seconds long. The cell phone was placed first on the top of a table, and then on the ground (Fig. 8.). In the semi-anechoic chamber, only the floor is reflective, other surfaces (the walls and the ceiling) absorb sound near perfectly. This way, when the source is on the top of the table, ground reflections are present, but when it is placed on the ground, as close to the reflective surface as possible, the difference between the distances the direct and reflected sounds must travel to the sensors is negligible. Thus, the impact of reflections on the source localization process can be investigated.

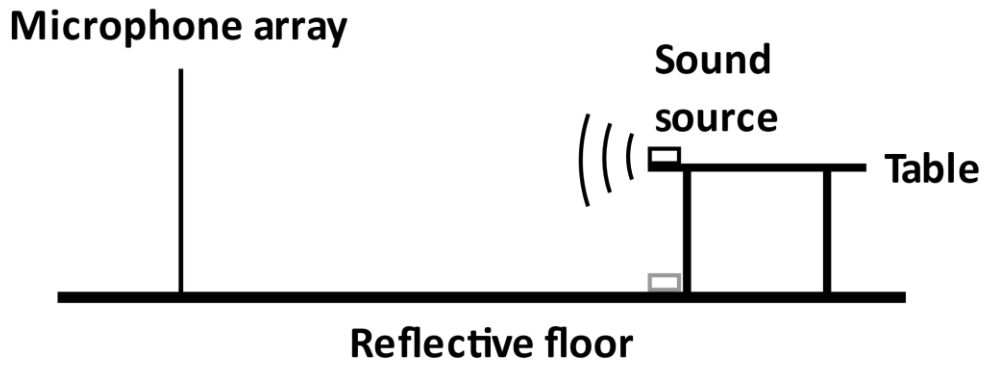


Fig. 8. Measurement setup in a semi-anechoic chamber to investigate the impact of ground reflections on the process of distance estimation.

Fig. 9. and Fig. 11. show the result of distance estimation, the former without, and the latter with ground reflections. The cell phone was approximately 5 meters from the plane of the microphones array (slightly less due to a small angular offset). When ground reflections are absent (Fig. 9.), both MUSIC and SAMV correctly estimates the distance. This is in contrast with the outdoor measurements because the environmental conditions inside the semi-anechoic chamber are close to ideal. Unfortunately, when ground reflections are present (Fig. 11.), neither algorithm can determine the distance. For the harmonic signal, the output of SAMV is highly inaccurate, around double the actual distance, while in the other three cases the results are similarly erratic to the outdoor measurements. However, SAMV still offers a substantial improvement over MUSIC because it can separate the direct and reflected sounds, and thus, it is capable of direction estimation in both cases (Fig. 10.).

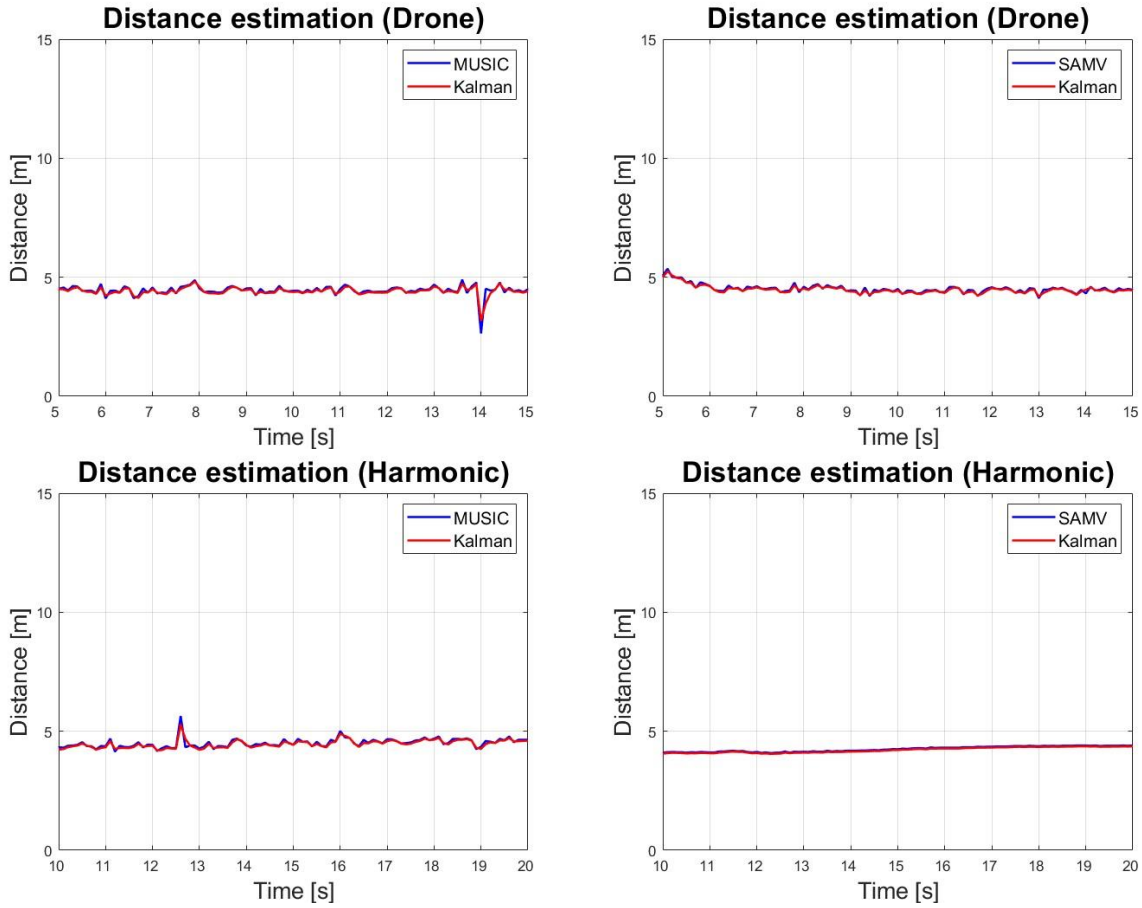


Fig. 9. Distance estimation during a measurement in a semi-anechoic chamber. The sound source was placed on the ground, so that ground reflections were absent.

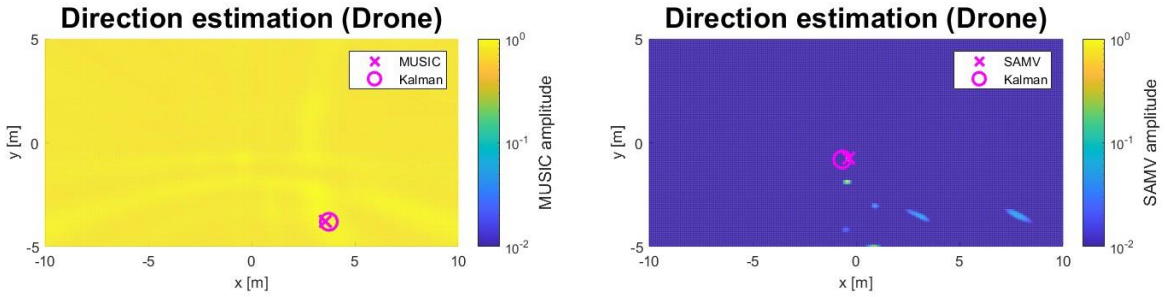


Fig. 10. Direction estimation during a measurement in a semi-anechoic chamber. The sound source was placed on top of a table, so that ground reflections were present.

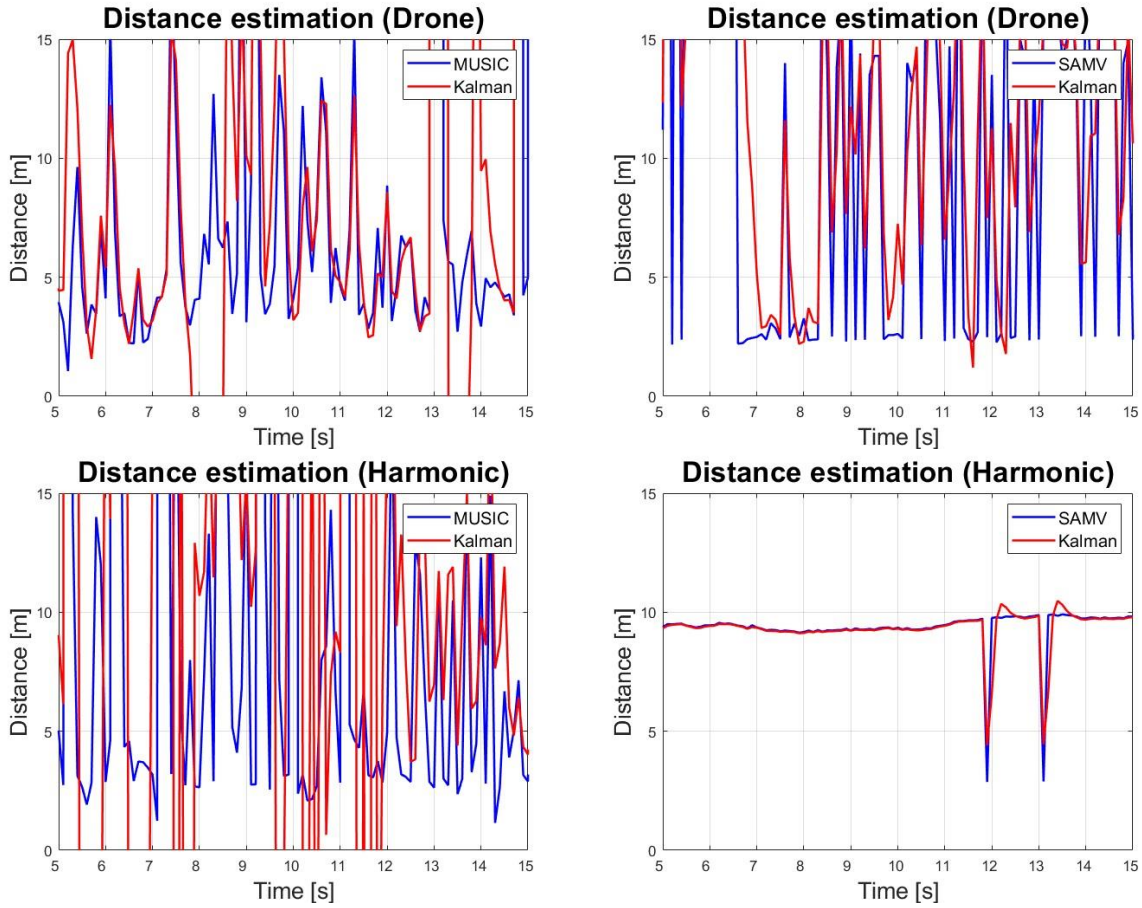


Fig. 11. Distance estimation during a measurement in a semi-anechoic chamber. The sound source was placed on top of a table, so that ground reflections were present.

3.5 Waveform and observed frequency

In the final simulation example, a point source is performing uniform motion parallel to the plane of the microphone array. The source emits a harmonic signal with an overtone at 1500 Hz. The central frequency of the observed narrow band changes between 1500 Hz and 1540 Hz. In Fig. 12., both the results of direction and distance estimation with MUSIC and Kalman-filter are depicted. The direction (the x coordinate) is temporally changing while the distance (the z coordinate) is constant. As expected, the farther the observed frequency is from the overtone at 1500 Hz, the worse the estimation becomes. The quality of distance estimation deteriorates faster than direction estimation, so it is more sensitive to correctly choosing the observed frequency. A consequence of this is that small temporal changes in the frequency spectrum might falsify distance estimation, while the algorithm can still determine the direction correctly.

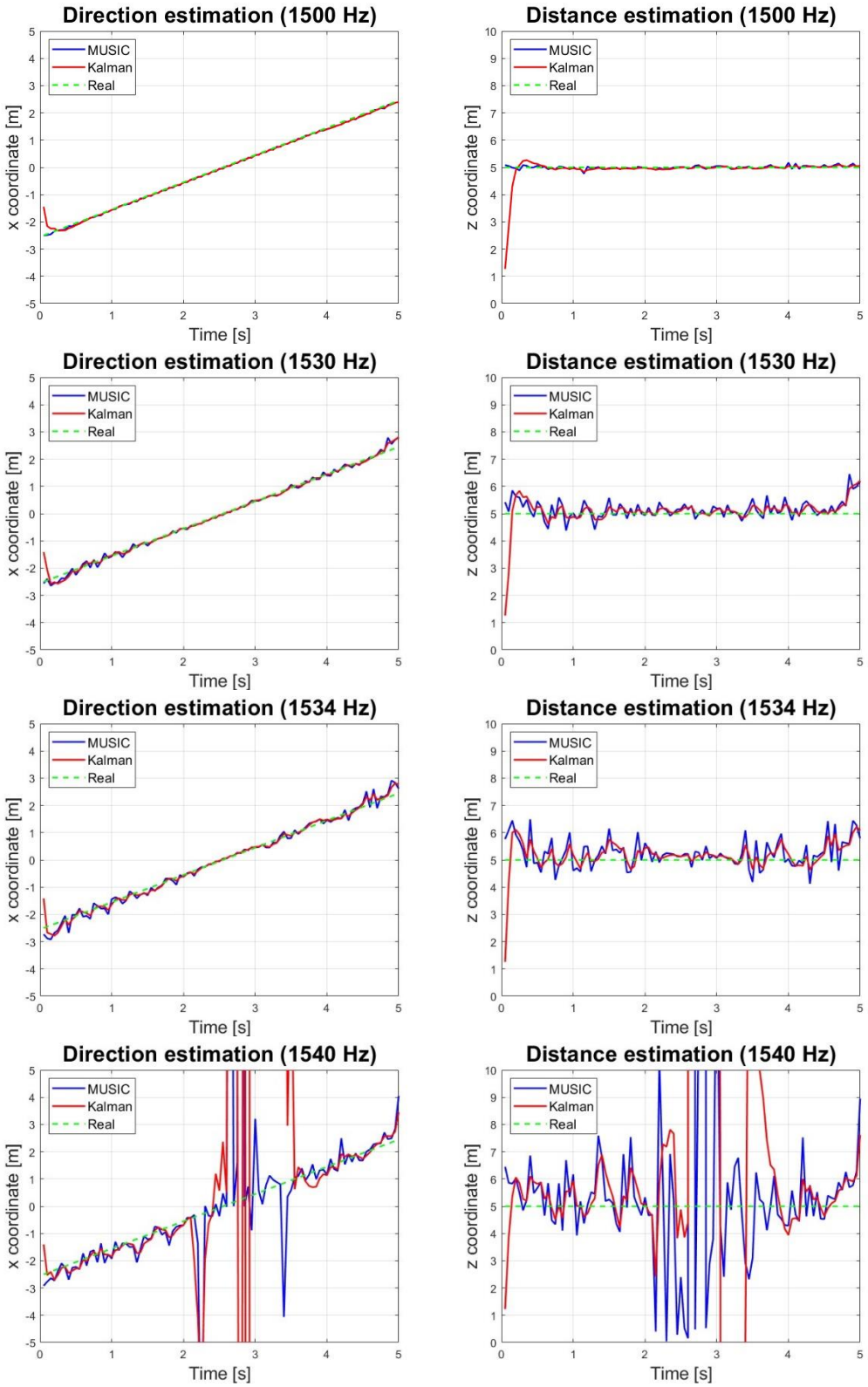


Fig. 12. Direction and distance estimation of a simulated sound source performing uniform motion. The source emits a harmonic signal with an overtone at 1500 Hz, and the central frequency of the observed narrow frequency band changes between 1500 Hz and 1540 Hz.

4 CONCLUSION

In this paper, we discussed the three-dimensional position estimation of sound sources with microphone arrays and beamforming algorithms. The beamforming algorithms used in our research are the MUSIC and SAMV algorithms, which we extended the Kalman-filter method to predictively track moving sound sources. The benefits of SAMV over MUSIC are its higher resolution and its ability to handle correlated signals, but its disadvantage is its higher computational cost. During preliminary simulations and outdoor measurements, the MUSIC algorithm was successful in determining the direction of the sound source, but its distance could only be determined during the simulations. There are many potentially critical factors that were neglected in the simulation that could cause this, and these factors need to be isolated and investigated. So far, we have investigated ground reflections, and the relation between the emitted sound and the observed frequency band. In a semi-anechoic chamber, when ground reflections were not present, both MUSIC and SAMV correctly determined both direction and distance. However, when reflections were present, only the SAMV algorithm achieved any success, and even then, it was only capable of direction estimation. In a simulation, where the observed frequency didn't perfectly align with the overtone of the emitted sound, we found that distance estimation is more sensitive to correctly choosing the frequency. These final results show two things: one, while ground reflections are indeed a critical condition that need to be accounted for in the 3D position estimation method, using the robust SAMV algorithm only achieves partial improvements; and two, slight temporal changes in the frequency spectrum also need to be accounted for by adaptively fitting the observed narrow band on peaks in the spectrum. Investigating other unfavourable factors in the future may also prove beneficial in making the 3D position estimation algorithm more robust.

ACKNOWLEDGEMENT

This work has been supported by the Hungarian National Research, Development, and Innovation Office under contract No. K–143436.

REFERENCES

[1] A. Xenaki, P. Gerstoft, K. Mosegaard: „Compressive beamforming". The Journal of the Acoustical Society of America, Vol. 136 (1), 2014, pp. 260-271, https://doi.org/10.1121/1.4883360.

[2] P. Gupta, S. P. Kar: „MUSIC and improved MUSIC algorithm to estimate direction of arrival". 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, 2015, pp. 0757-0761, https://doi.org/10.1109/ICCSP.2015.7322593.

[3] L. Yaning, F. Juntao, R. Xinghao, M. Le: „An improved MUSIC algorithm for DOA estimation of non-coherent signals with planar array". J. Phys.: Conf. Ser. 1060 012026, 2018, https://doi.org/10.1088/1742-6596/1060/1/012026.

[4] H. Abeida, Q. Zhang, J. Li, N. Merabtine: „Iterative Sparse Asymptotic Minimum Variance Based Approaches for Array Processing". IEEE Transactions on Signal Processing, Vol. 61 (4), 2013, pp. 933-944, https://doi.org/10.1109/TSP.2012.2231676.

[5] X. Zhang, J. Sun, X. Cao: „Robust direction-of-arrival estimation based on sparse asymptotic minimum variance“. The Journal of Engineering, Vol. 2019 (21), 2019, pp. 7815-7821, https://doi.org/10.1049/joe.2019.0720

[6] Y. Cai, X. Liu, Y. Xiong, X. Wu: "Three-Dimensional Sound Field Reconstruction and Sound Power Estimation by Stereo Vision and Beamforming Technology". Applied Sciences, 2021, 11(1), 92, https://doi.org/10.3390/app11010092.

[7] J.-M. Valin, F. Michaud, J. Rouat: "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering". 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 2006, pp. IV(841)-IV(844), https://doi.org/10.1109/ICASSP.2006.1661100.

[8] R. Merino-Martínez, B. von de Hoff, D. Morata, M. Snellen: "Three-dimensional acoustic imaging using asynchronous microphone-array measurements". 9th Berlin Beamforming Conference 2022, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S08.pdf.

[9] E. Sarradj: "Three-dimensional gridless source mapping using a signal subspace approach". 9th Berlin Beamforming Conference 2022, https://www.bebec.eu/fileadmin/bebec/downloads/bebec-2022/papers/BeBeC-2022-S06.pdf.

[10] M. U. Liaquat, H. S. Munawar, A. Rahman, Z. Qadir, A. Z. Kouzani, M. A. P. Mahmud: "Sound localization for ad-hoc microphone arrays". Energies 2021, 14(12), 3446, https://doi.org/10.3390/en14123446.

[11] J. Novoa, R. Mahu, A. Díaz, J. Wuth, R. Stern, N. B. Yoma: "Weighted delay-and-sum beamforming guided by visual tracking for human-robot interaction". 2019, https://doi.org/10.48550/arXiv.1906.07298.

[12] R. Schmidt: "Multiple emitter location and signal parameter estimation". IEEE Transactions on Antennas and Propagation Vol. 34, 1986, pp. 276–280, https://doi.org/10.1109/TAP.1986.1143830.

[13] M. Mohanna, M. L. Rabeh, E. M. Zieur, S. Hekala: "Optimization of MUSIC algorithm for angle of arrival estimation in wireless communications". NRIAG Journal of Astronomy and Geophysics, Vol. 2 (1), June 2013, pp. 116-124, https://doi.org/10.1016/j.nrjag.2013.06.014.

[14] Q. Zhao, W. Liang: "A Modified MUSIC Algorithm Based on Eigen Space". In: Jin D., Lin S. (eds) Advances in Computer Science, Intelligent System and Environment. Advances in Intelligent and Soft Computing, Vol 104. Springer, Berlin, Heidelberg, 2011, https://doi.org/10.1007/978-3-642-23777-5_45.

[15] D. Simon: "Optimal State Estimation – Kalman, H∞, and Nonlinear Approaches". John Wiley & Sons, Inc., Hoboken, New Jersey (2006).

[16] Z. Belső, B. Gáti, I. Koller, P. Rucz, A. Turóczi: "Design of a nonlinear state estimator for navigation of autonomous aerial vehicles". Repüléstudományi közlemények (Aviation scientific publications) XXVII/3 pp. 255–276 (2015).